Paweł Marczewski
MIMUW

Problemy świadomych maszyn i ich relacje z ludźmi

Wobec ciągłego wzrostu mocy obliczeniowej komputerów powstaje pytanie - czy wreszcie uda się zrealizować odwieczne marzenie ludzkości i stworzyć myślącą istotę? I jakie będą tego konsekwencje?

Pochodzenie AI

To, że świadomy program zostanie stworzony przez człowieka, nie znaczy jeszcze, że człowiek będzie rozumiał dokładnie jego działanie. Istnieje wiele rozwiązań polegających na generowaniu układu w dużym stopniu automatycznie (np. algorytmy genetyczne, sztuczne sieci neuronowe), określając tylko warunki docelowe i kierunek rozwoju. Biorąc pod uwagę złożoność zadania, może się okazać, że większość stworzonego programu będzie napisana przez komputer, a nie przez nas.

Rozwinięciem tej idei jest "seed AI" - algorytm w pewien sposób "rozumiejący" własną budowę i zdolny do stopniowego ulepszania samego siebie. Doprowadziłoby to do "eksplozji inteligencji", szybkiego i potencjalnie nieograniczonego rozwoju. Skonstruowanie takiego systemu pozostaje jednak w fazie marzeń (istnieją jednak organizacje, które dążą do jego powstania).

Internet

Internet wydaje się najbardziej naturalnym "środowiskiem życia" dla sztucznej inteligencji. Zauważmy jednak, że obecnie maszyny wcale nie mają w nim takich samych praw jak ludzie. Zalew spamu zmusił wiele stron do założenia CAPTCHA ("Completely Automated Public Turing test to tell Computers and Humans Apart") - testów może nie mających wiele wspólnego z oryginalnym testem Turinga, ale równie skutecznie odróżniających maszynę od człowieka.

Trudno powiedzieć, jak rozwinie się CAPTCHA, jeśli maszyny będą stawać się coraz bardziej inteligentne. Z jednej strony trudno powstrzymać program o ludzkiej inteligencji, zdolny dobrze naśladować człowieka, z drugiej mając do dyspozycji automat o takiej inteligencji łatwiej jest odsiewać sensowne wiadomości od niechcianych reklam.

Rozwiązałoby to problem spamu, ale nie odróżnienia człowieka od maszyny - możliwe więc, że aby to zapewnić, część stron zrezygnuje z internetowej anonimowości i każe rejestrować się przy pomocy własnych danych osobowych. Skuteczna realizacja spowodowałaby podział internetu na części "dla wszystkich" i "tylko dla ludzi".

Nauka

Gdy w 1976 roku udowodniono z pomocą komputera twierdzenie o czterech barwach, nie zyskało to uznania ogółu matematyków - uznano, że dowód zbyt skomplikowany, aby człowiek go zrozumiał, jest "brzydki". Możliwe jednak, że komputeryzacja jest po prostu istotną zmianą w historii matematyki (znany informatyk Edsger W. Dijkstra twierdzi wręcz, że jest to pierwsza istotna zmiana: przejście od

ograniczonych możliwości obliczeniowych ludzkiego mózgu do ciągle rosnących możliwości komputerów).

Wzrost mocy obliczeniowej, szczególnie w połączeniu z ewentualnym potencjałem twórczym AI, może też znacznie przyspieszyć rozwój innych nauk. Według zwolenników Seed AI, "eksplozja inteligencji" może nieść ze sobą eksplozję postępu technicznego, nazywaną Singularity.

Kontrola

Jednym z ulubionych motywów fantastyki naukowej jest bunt maszyn - rozmaite scenariusze, w których myślące maszyny wymykają się spod kontroli, próbują przejąć władzę albo wręcz zniszczyć ludzkość. Może to być po prostu przejaw antropocentryzmu i nadawania przyszłym AI zbyt wielu cech ludzkich, warto jednak problem dokładnie przeanalizować.

Pozwolenie (nie tylko myślącym) programom na kontrolę nad różnymi aspektami naszego życia - jak kierowanie pociągami, samolotami, czy choćby mycie szyb i nalewanie kawy - słowem, wszystko, co pozwala im na fizyczną ingerencję - budzi duże obawy. I nie chodzi tu tylko o słynny "bunt maszyn", ale też o zwykłe wątpliwości, czy tak skomplikowany układ zawsze będzie robił to, co powinien.

Możliwości AI można ograniczyć po prostu nie dając jej takiej kontroli, i pozostawiając wszelkie istotne aspekty ręcznie zaprogramowanym, odizolowanym maszynom. Może się jednak okazać, że prędzej czy później sztuczna inteligencja przejmie kontrolę (wskutek działań własnych lub człowieka); powierzenie kontroli AI jest też kuszące choćby z uwagi na to, że programowanie komputera jest dla człowieka zadaniem trudnym, a błędy zdarzają się bardzo często. Koncepcja Singularity nie wyklucza wręcz, że AI uzyska nieograniczone możliwości fizyczne - warto by się więc do takiej sytuacji dobrze przygotować.

Idealnie byłoby, gdyby programowi można było podać dobrze określone cele, których nie może w żaden sposób obejść. Przejawem tego, bądź co bądź, marzenia są trzy prawa robotyki sformułowane przez Isaaca Asimova. Sprowadzają się one do następujących trzech priorytetów: maszyna nie może krzywdzić ludzi, musi wykonywać ich rozkazy oraz chronić własne istnienie. Jest to jednak spore uproszczenie i jako takie łatwe do obejścia (choćby dlatego, że pozostawia pewną swobodę interpretacji). Budzić zastrzeżenia może również np. to, że niemożliwe jest zastosowanie takich robotów na wojnie.

Eliezer Yudkovsky, twórca koncepcji "seed AI", zaproponował bardziej fundamentalne pojęcie - "friendliness", przyjaznej sztucznej inteligencji: polegałoby to na wbudowaniu przyjaznego nastawienia do człowieka tak głęboko, aby nie ulegało ono zmianie podczas ewolucji układu. Zagrożeniem jest bowiem nie tylko wrogość AI wobec ludzkości, ale także zwykła obojętność wobec jej losu. Wymaganiami stawianymi "friendly AI" jest:

- szeroko pojęta "życzliwość" wobec ludzi
- zachowywanie i przekazywanie tego nastawienia stworzonym przez siebie systemom
- inteligentna interpretacja nakazów
- dążenie do samodoskonalenia
- "przewaga pierwszego ruchu" - a więc przeciwdziałanie powstaniu innych AI, nie przestrzegających powyższych zasad).

Bez wątpienia są to cele trudniejsze do realizacji od trzech praw robotyki. Nie ułatwia również fakt, że należy się liczyć zarówno z podatnym na uszkodzenia sprzętem, jak i nieuchronnymi błędami ludzkich projektantów.

Do stworzenia prawdziwej sztucznej inteligencji jesteśmy prawdopodobnie bardzo daleko; wielu twierdzi wręcz, że nigdy się to nie uda. Warto jednak zastanowić się nad takim scenariuszem, gdyż może on prowadzić do najważniejszej zmiany w historii...

ŹRÓDŁA

- Wikipedia (Technological singularity, Friendly AI, Seed AI, Ray Kurzweil)
- On the cruelty of really teaching computing science, E.W.Dijkstra, 1989
- Creating Friendly AI, Singularity Institute, 2001